Variables, designs, hypotheses

LECTURE 2

Scientific Methods

- experiment:
  - vary some i.v. while holding everything else constant
  - measure changes in some d.v. changes in dv must be caused by changes in i.v.
- quasi-experiment:
  - same ↑, but i.v. can't be manipulated (eg. gender differences)
  - potential probs with confounding (extraneous) variable
- correlational design:
  - no manipulation
  - measure 2 or more variables, determine the extent to which they relate together
  - cannot infer casuality

Design

- between-subjects
  - each condition is applied to a <u>different</u> group of subjects.
  - often the only available option: eg.
    - measuring differences btw men/women
    - measuring ability to learn a particular tasks using two different methods. (once learned, cannot be unlearned).
  - individual differences
    - by assigning subjects randomly to groups, individual differences should roughly balance between groups.
- within-subjects
  - same subjects performs at all levels of the independent variable.
  - also known as repeated measures designs: subjects repeat the measure for each different condition.
  - This design is generally much more powerful: each subject is their own control so individual differences are ruled out.

Variables

- independent
  - 1 or more
  - each can have 2 or more levels

eg. 1 i.v = time of day
- before meal
- after meal
- 3h after meal

3 levels

eg. amount of coffee
- 1 cup
- 2 cups

2 levels

- dependent
  - 1 or more
  - what we actually measure
  - Does A'level maths → effect on undergraduate Psych degree?

d.vs = 1st year exam marks, 3rd year project

- confounding & nuisance →
  - causes confusion in the interpretation of results
  - this occurs when some aspects of the experiment varies systematically w/ i.v.
  - aim: eliminate any potential confounding variables.

exam
- the [illegible] stimulus was more interesting than verbal stimulus
- "heavy drinker" group younger than "light drinker" group

- introduces noise but do not <u>systematically</u> bias the results

examples:
- occasional loud noises during the experiment.
- some subjects are older than others.

- several types/measure scales

eg. nominal scales (type of beer) – categories, mutually exclusive
- ordinal scales (quality of each beer, 1-5) – ranks, position meaningful
- interval scales (mean temperature of beer) – difference btw values is meaningful
- ratio scales (beers/night) – same as interval, 0 = meaningful != nothing "twice as big"

SHOUT
SHARE
IT

within-subjects

ADVANTAGES

- each subject is their own control

DISADVANTAGES

- designs may lead to a confound with order effects.
  can be minimized by "counterbalancing"
  ↳ eg. in an experiment with two levels of i.v.
  - half subjects recieve level 1 first
  - half subjects recieve level 2 first

- matching-subjects
  - eg. pair students in classes by their IQ.
- statistically, data are treated like a regular within-subjects, with a pair of subjects being treated as 1 individual tested on 2 levels.

Q.72 What is not an advantage of repeated measures in comparison to between-group design?

→ Researchers can study cross-cultural effects more easily.

advantages

→ each participant is his own control
→ more powerful
→ requires fewer participants overall

Q75 counterbalancing is used:

a) to overcome systematic bias caused by boredom effects
b) overcome systematic bias caused by pratice effects

Hypotheses

→ We reject the null hypothesis when the probability of null hypothesis being true (p) is less than some critirion (α).

- if probability $p < 0.05$ we mean that there was less than 1 in 20 probability of it happening by chance
- p is the probability of collecting this data assuming the null hypothesis to be true.
- α is the critirion level we set: if $p > \alpha$ we failed to reject the Null Hypothesis

This doesn't mean that this is true: just that we have not got any real reason to reject it yet.

- experimental hypotheses
  questions we want to adress based on general knowledge
- statistical hypotheses
  precise statements about the data to be collected
  - Null Hypotheses H0
    states that the differents samples we look at come from the same population
    → H0: all the means are equal
    → H0: all the distributions are the same
    - it is our default. we accept it unless we have good evidence to the contrary.
  - Alternative hypothesis H1
    - the logical opposite of the null hypothesis.
      ← mutually exclussive w/ H0
      ← exhaustive (they cover all possible outcomes in the experiment).

Descriptive Statistics

LECTURE 3

Frequency Distributions →

- categorical data – frequencies, percentages
  - few possible score
  - many possible scores (frequency ranges).
- discrete data
  - cummulative
  - percentage frequency
  - cummulative percentages

frequency ranges
- too many scores
- create ranges

discrete variables
have certain fixed values
eg. N° of objects
eg. ordinal scale variables

continuous variables
can take any fractional value
eg. amount of time
eg. interval/ratio scales

Central tendancy

- We want to sum up the data, using only the important, main point

- mode – most common
  → sometimes more than 1 value –
  → varries depending on bin size –
  → used for categorical data + (nothing else can)
  → always gives a real data value +
  → generally used for (nominal) data only
  → ranges distribution: modal range = 496–568 group

- median – middle
  ↳ set data
  → insensitive to outlying data +
  → often gives a real data value +
  → intuitively appealing +
  → ignores a lot of the data –
  → difficult to calculate w/out computer –
  → best measure for (ordinal) data sometimes for skewed interval or ratio data

- mean – sum divided over number of samples

$$\bar{X} = \frac{\Sigma x}{n}$$

  ↳ uses all data (powerful) +
  → very sensitive to outliers/skew –
  → not always giving a meaningful value (2.4 children?) –
  → only meaningful for (ratio) and (interval) data –
  → best measure for (normally distributed data)

— central tendency
— spread of scores
] need these when: eg scored 62% is it good or bad relative to rest.

✡ Measures of spread are related to measures of central tendency:
- mode: no associative measure
- median: associated with range or IQR
- mean: associated with variance or standard deviation

Range
highiest value − lowest
− very sensitive to outliers.

IQR : $Q_3 - Q_1$
ignores most extreme values
Q_1: score achieved by the lowest 25% of the data. $N \times 25\%$.
Q_2: median
Q_3: 75% of data
eg. $N = 163$
$Q_1 = 163 \times 25\% = 42.5 \rightarrow 43^{rd}$ score
$= 488.1$ cm
+ → less sensitive to outliers
+ → less assumptions about data
− → hard to calculate on hand
− → not all data used.

Variance σ^2
$$\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{N}$$
how far a set of scores are spread out from their mean

eg.

score/15	score − mean	(score − mean)^2
2	−5	25
4	−3	9
6	−1	1

sum = 35
mean = 2
$\sigma^2 = \frac{35}{2} = 17.5$

+ → uses all data
+ → forms the basis of other tests
− → more sensitive to outliers
− → requires a normal distribution
− → does not have sensible unit (ms^2??)

Standard Deviation σ
− variance is useful but would be more useful if the units were sensible
− square root of variance
− same unit as the one variable uses
− $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$

from sample $= s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N-1}}$
why? we get a better unbiased estimate of what the population standard deviation if we could only measure the sample's standard deviation.

* When data are skewed:
chose median & IQR (less sensitive to outliers)

* When data are not skewed:
choose mean & standard deviation)
(sensitive to outliers)

Probability, Sampling & Distributions

LECTURE 4

probability Theory

→ often in statistics, we wish to compare our results with how likely they are to occur by chance.

→ Interested in the study of random phenomena. eg. chance

Normal Distribution

- Gaussian Equation

→ The equation to find the height of the distribution (y) at any point on

$$y = \text{peak} \times e^{-\frac{(x-\mu)}{2\sigma^2}}$$

x = desired score

y = Frequency at that x

μ = mean of pop

σ = standard deviation of pop

→ The peak of the distribution can be easily calculated:

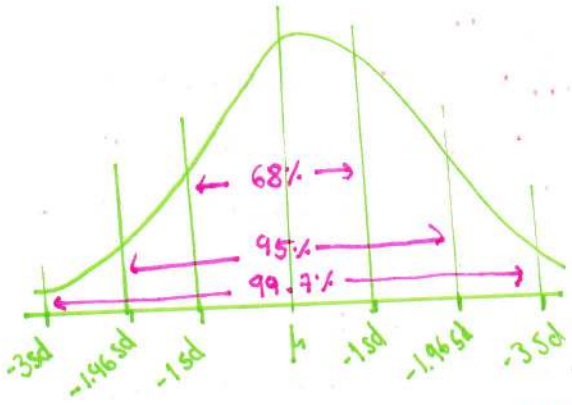
$$\text{peak} = \frac{N}{\sigma\sqrt{2\pi}}$$

N = number of total

σ = standard deviation of pop

→ From the (mean) and (standard deviation) we can predict the value of N for any value of x.

Standard Deviation



if for this data mean = 43.5 ms
std = 8.0 ms:

Parametric Tests

→ many tests are based on estimates of population parameters such as mean, σ

→ parametric tests assume that the mean and standard deviation accurately represent the population distribution

- Transforming data
  operations (mathematical) to all scores recorded
- This may help reduce the impact of outliers
- Z-scores

eg. stats exam: mean = 55
σ = 15
is 68 good? Calculate z-score:

$$z = \frac{\text{score} - \bar{x}}{s}$$

It tells us how many σ we have ↑ or ↓ the mean

Skewness

- measuring skewness

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

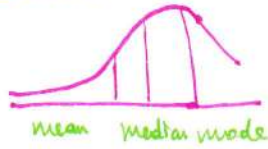
↳ remove influence of variability

- unskewed

← mean < median < mode

- positive skewed

mean > median > mode



- negative skewed

Sampling Error

mean of sample surveys differs from a population

but: the mean of each sample differs from the true mean (and from other samples)
= Sampling error

Standard Error of Mean

tells us how much a statistic is likely to vary from one sampling to another.

S.e.m. = effectively a measure of how confident we are that we know the true population mean.

depends on:
- variability of original data (σ)
- number of data used to create the sample mean.

$$s.e.m. = \frac{\hat{s}}{\sqrt{N}}$$

– ↑ people, ↑ representative

Standard Deviation Vs s.e.m.

- σ : the amount that scores vary
- s.e.m. : measure of our confidence that know the true population mean

example : repeatedly throwing darts at a board

– s.d. measures how much darts vary around your target number (i.e. how good player you are?)

– s.e.m. measures how confident we are that we know our target number (i.e. dependent on how good we are and how many times we throw).

Confidence Intervals

mean is useful but doesn't have a very intuitive ring to it

→ Where would you find the mean if you would run an experiment and collect 95%

how? just as 1.96 times the s.d. gives the 95% of all scores, 1.95 times the sem gives 95% of the sample means.

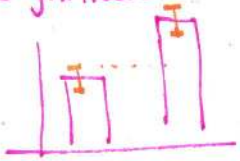
eg

$$c.i._{95} = \bar{X} \pm 1.96 \cdot s.e.m.$$

Error Bars

- non-overlapping error bars imply significant differences

eg.



Using SPSS

LECTURE 5

rows represent subjects
columns represent variables

- Summarising Data

number of participants
male female ratio

→ Analyze / Descriptive Statistics / Descriptives

↳ options: (mean, st. dev., max.)

→ can also add frequencies

example: st.dev comes to be 40.
This is a very large number ∴
→ OUTLIERS

Dealing with Outliers

= conduct a test on only the middle range of data?

= Transform data to get rid of the skewness.

Graphs

selecting the variables to plot:

Varriables): i) Number of words for Our F
ii) Number of words for Another

Category Axis = i) Gender

- selecting subjects of data.

eg. just younger subjects, just the ones who filled question 4.

↝ Some "if" condition

click on "selecting Cases"

Top - Right

Comparing Means Part I

... are they so different that we think the i.v. truly has an effect on the d.v.? Is it statistically significant?

- reliable
if we repeat the experiment with a similar but not identical group subjects, will we observe the same result?

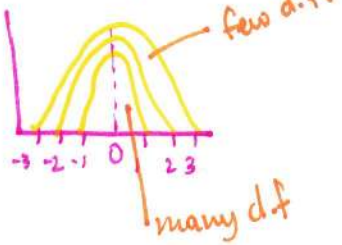
t-test

- 3 main versions
  - independent samples test
  - paired-samples test
  - one-sample test
- t-statistic
is a value depends on:
  - the size of the difference between 2 conditions
  - the s.e.m. of that difference (which in turn depends on the variability of the data & the number of measures we take).

Once we have calculated t, we can look up the associated p value in a table.

* We need to know the number of degrees of freedom.
(N guests are coming for dinner. How many can chose their own seat? N-1)

Degrees of freedom change the probability distribution of the test statistics.



- related to number of subjects
- relationship is different for each different test (eg. the choice of seats depends on seating method)
- df is needed to work out p from t

Independent-Samples t-test

LECTURE 6.

- for between-group designs.
- also known as:
  - unpaired means t-test
  - unrelated samples t-test
  - between-groups t-test
  - independent-measures t-test
  - independent two-sampled

Based on:
- size of the difference between conditions
- the s.e.m. of that difference (which is the confidence we have that our mean is the true one)
- equation

① Calculating mean of differences.
(mean of GP1) − (mean of GP2)

② Calculating standard error

$$s.e.m.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

(X) t is the mean difference divided by the standard error difference

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

- t-test based on mean & standard deviation

assumes that these adequately describe the data:

① Data must be on a ratio/interval scale

② Data must be normally distributed

③ Data on 1 group must be independent of data in other

④ Data in the two groups must have similar variance.

- SPSS

example: Do people chose to sit closer to the front if they need optical correction?

....

- two student groups i) near front
  ii) near back
- measure/ask for optical correction

analysis: between-subject t-test

- i.v. distance from screen
- d.v. dioptres of optical correction

→ Analyse/Compare Means/Independent Samples ...

test variable: optical correction

grouping variable: position in lecture room

(define groups as GP1 = 0 (near)
GP2 = 1 (far)

- output

① You have to chose the appropriate row:

	F	Sig	t	df	Sig (2 tailed)
Equal Variances assumed	1.533	.241	-.808	11	.436
Equal Variances not-assumed			-.843	9.688	.419

→ if sig > 0.05 then variances are equal
use upper row

→ if sig < 0.05 then variances are not equal
use lower row

② Obtain the test statistic (t) the degrees of freedom and value of p (column labeled "Sig (2-tailed)")

- reporting.
- p > 0.05 → there was no significant difference
- p < 0.05 → there was a significant difference

overall

- Is our observation unlikely to have occured by chance?

→ we want to know if p < alpha

- we need to calculate p (alpha is fixed)
  → calculate some test-statistic eg(t)
  → from the test-statistic and d.f, calculate p.

(t → d.f → p)

Q.35 How do you compensate for heterogeneity of variance in the independent samples-t-test?

→ correct the degrees of freedom

Comparing Means Part II

LECTURE 7

Paired-Means t-test

- for within-subject designs
- also known as : related-sample t-test
  within-subject t-test
  repeated measures
- similar to the between-subject t-test but uses subjects as their own control
  → more powerful, more likely to find an effect.
- equation
  - Size of effect divided by the standard error of the effect
- Mean of the differences for the pairs divided by the standard error of the differences between pairs.

$$t = \frac{\Sigma (X_1 - X_2)/N}{S.e.m._{(X_1 - X_2)}}$$

example: Navon Task

measuring reaction time of subjects trying to detect the presence of a large letter compromised of:

- similar (consistent) letters
- disimilar (conflicting) letters

→ we want to know whether subjects are significantly different in their reaction times in the two conditions.

gender / age / RT conflicting / RT consistent

→ display normal curve
Graphs / Legacy Dialogs / Histogram

- ~~output~~ Analysis-t-test
  Analyze / Compare Means / Paired Samples t-test
  variable of interest : Conflict RT - ConsistRT

Only 1 column:
correlation = indicates if subjects were acting consistently across different conditions

→ gives info about the size of difference between the conditions.
(Conflict RT - Consist RT)

- SPSS
  - same steps as for the independent samples t-test
  1. enter data in data view
     one subject on each line
     one i.v. level in each column
  2. Save → run test

.....

one column corresponds to:

- one between-subject variable: one column will be your i.v. or "grouping variable" the other is your d.v.

	Coffee Dose	RT
1	0	300
2	0	450
3	0	320
4	1.0	410
5	1.0	280
6	2.0	350

between-subject

within-subject

- one Level of a within-subject variable: no column for your i.v., one column for d.v. at each level of your dv.

	RT_1Cup	RT_2Cups	RT_3Cups
1	300	280	280
2	450	350	300
3	320	300	270
4	410	290	420

Correlations

correlation Vs Causality

- B could depend on A as much as A depends on B
- A and B could depend on some other ~~joined~~ joint variable

- correlations are useful
  - highlight possibilities
  - null effects

- design

measure 2 or more variables

→ null hypothesis:

no relationship between variable.

→ Alternative hypothesis:

there is a relationship between variables

- measures of association
  - covariance (parametric)
  - Pearson's coefficient of Correlation (parametric)
  - variance explained (variance accounted for)

Covarience

- the mean of the product of the deviations from the mean
- how much 'random variables change together'

$$cov = \frac{\Sigma(\bar{x} - x)(\bar{y} - y)}{N}$$

* if we want to express covariance in a standard unit of measurement simply divide by the σ.

* eg: look @ book

strong relationship

X	Y	Prod of devs
1	3	1x1 = 1
2	4	0x0 = 0
3	5	-1x-1 = 1

cov = 2/3 = 0.667

no relationsp

X	Y	Prod of Devs
1	2	1x0 = 0
2	2	0x0 = 0
3	2	-1x0 = 0

Cov = 0/3 = 0

- Cons: covariance is related to variance. when x and y vary a lot, the covariance is large.
- pros:
  - when X & Y go up together, the covariance is large and positive
  - when X goes up when Y goes down, the covariance is large and negative
  - when X and Y have no relation, covariance is zero

Pearson's Coefficient of correlation

LECTURE 8

- based on covariance
- by dividing by the standard deviations of the variables, it is independent of the overall variability.

$$r = \frac{cov(x,y)}{S_x S_y}$$

where S_x and S_y are the s.d's of the samples x and y

properties:

- only dependent on how well-related the variables are, not how much they vary
- Takes values between -1 and +1
  - * -1 = when X goes up, Y goes down
  - * +1 = means when X goes up, Y goes up
  - * 0 = no linear relationship btw X & Y

r can also be used for experimental data eg. Sternberg Experiment

facts @ 37

- can be used as an effect size measure
- varies from -1 - 1
- can be used on ranked data
- can be used for experimental designs.

measures of association

- Sperman's Rho
  - coefficient of correlation related to Pearson's r.
  - non-parametric: doesn't assume data are equally distributed
  - use it for ordinal data Q.42
  - works by:
    1. Ranking data
    2. Working out Pearson's r on the ranks
- SPSS

Q = how do students' reaction times correlate in different conditions of the Navon Task?

Graphs / Legacy Dialogs / Scatter / Dot

→ computing the coefficients

Analyze / Correlate / Bivariate

- multiple pair-wise correlations

→ (run correlations on more than one variable at once)

Variance Accounted for

↳ another way to think of correlations

↳ expressed as a percentage or a fraction

↳ simply calculated by (*) squaring Pearson's r

eg. $r^2 = 0.876^2 = 0.767$

~ 77% of the variance in the consistent condition can be accounted for by the variance in the conflicting condition.

(*) Q.39. How much variance can be explained by a correlation of 0.9? → 81%.

partial correlations

sometimes the correlation between to variables A and B can be explained by variable C.

A = times one gets drunk

B = exam score

C = drinking → not at home studying

* A partial correlation measures the part of the correlation that is not related to variable C

Terminology

- Variable C → 'partialled out' / 'held constant'
- a simple correlation is called a zero-order correlation
- a first order (partial) correlation is one that partials out a single variable constant
- a second order (partial) correlation holds 2 variables constant

Select / Analyze / Correlate / Partial

↳ output suggests:

"... after accounting for age, we still have a very strong correlation between reaction times. (in fact, partialling out age had no effect on data — irrelevant).

Q40. The correlation of a variance with itself is:

→ always +1

Issues with Significance Testing

what is power?
probability of finding an effect given that one trully exists.

p = probability of obseving data given that H0 true power (denoted 1-β) = probability of finding p < alpha given that H0 is false.

POWER & ALPHA

	H0	H1 → real world
H0	true negative	type II error ('miss') probability = β
H1	type I error ('false alarm') probability = α	true positive ('hit') probability = 1-β

↓ our sample

Q4: There is meaningful effect in the population, when in fact there is not

Factors Affecting Power

power, 1-β = probability of finding an effect, given that there is actually one.

- power will obviously be governed by:
  - size of effect
  - number of subjects
  - choice of alpha
- also (less obviously)
  - sources of variability
  - experimental design
  - choice of test
- tails of a test

SIZE OF EFFECT

metal Vs wooden ball falling
→ metal falls faster, ↑ chance of measuring the difference in their speeds if its large.

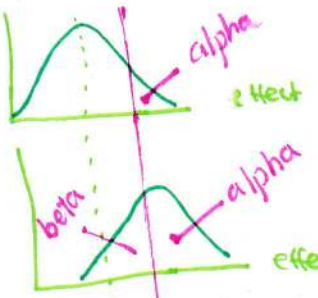
Power: ① The ability of a test to detect an effect.
Q.6
② Use power to determine how big a sample is required to detect an effect of a certain size.
③ Power is linked to the probability of making type I error

NUMBER OF SUBJECTS | LECTURE 9

↑ subjects, more chance the effect is real.

- ↑ subjects → real means observed
- ∴ greater confidence that we can spot the difference.

CHOICE OF ALPHA



if H0 were true

if H1 were true

* choice of alpha depends on the different outcomes of false positives or false negatives

- if the outcome of falsely rejecting the null hypothesis is bad, then use a low alpha (high criterion).
  → working if wheat makes one grumpy a set it to 0.001.
- if the outcome of falsely accepting the null hypothesis is bad then use a high alpha (low criterion)
  → taking a drug checking for severe side-effect (eg. death) one would pick a large α = 0.1

In Psychology, α = 0.05 standard.

SOURCES OF VARIABILITY

minimize noise variables.
eg RT = time to identify emotion
= time to respond (click, read)

therefore, present the faces for a variable length of time and see how long is enough for the perception to be accurate.

→ additional noise is removed

EXPERIMENTAL DESIGN

where possible, try to use a within-subject design (and then counterbalance to avoid order effects)
Due to individual differences which is a source of variability

CHOICE OF TEST

not only about between-within
• Tests that take into account the size of differences between scores are generally more powerful.

group 1	group 2	group 3
0.8	0.5	-0.3
0.5	1.2	0.7
1	0.9	-0.1
0.9	1.5	0.6
0.7	1.4	0.7
0.7	0.6	-0.1
0.5	1.1	0.6
0.6	1.5	0.9

→ gp 1 scores larger than gp 2, half the times
→ gp 2 scores significantly larger overall, using an unpaired samples t-test.

TAILS OF TEST

• two-tailed test
we hypothesise that subjects' scores in condition A will be different than there in B.
- The p-value is the probability of seeing this big difference in either direction.
- harder to find an effect

• one tailed test
we hypothesise that subjects' scores in condition A will be greater than those in B
- the p-value is the probability of seeing this is a big difference in a given direction.
- better power when there is a clear prediction in one direction

→ 1-tailed tests are more powerful than 2-tailed. More likely to have a positive result for the same data.
→ 1-tailed tests are based on some prior knowledge or hypothesis
→ if you predict an effect in 1 direction and then find one in the other, with a 1-tailed test you conclude "no difference found"

Calculating alpha on SPSS

required: size of effect
number of subjects
choice of alpha
OR SPSS calculates the sample size you need to give the power that you're looking for.

* family-wise error
→ is an inflated Type I error.
If we run N experiments, or N tests in a single experiment, roughly how many tests will give false positive results?

Significance Vs Importance

- what if effect is really there but so small that it doesn't mean anything?
- what if d.v. doesn't really measure what we want it to measure?

Eg. "Men significantly more intelligent"
→ maybe gender caused a slight difference but is not a big predictor of intelligence.
→ Effect may be small
→ d.v. may not be measuring the right thing (IQ tests)

Categorical Data

LECTURE 10

χ^2 - chi square

(Q35) Chi-Square is a test of association

- Is the test to determine how often observations fall into a particular category. (categorical data.)

- χ^2 compares how many observations fall into a category with how many were expected based on some model.

- χ^2 by hand

one i.v. eg. test whether students choose randomly what career/category to go to

- Null: the different frequencies observed for different categories were expected by chance

- Alternative: the different frequencies observed for different categories reflect actual differences between the categories.

4 STEP APPROACH:

① Create a 'contigency table' to record the frequency of each category. Categories must be mutually exclussive & exhaustive

	Male	Female	Total
Dog	42	10	52
Cat	52	9	61
Total			

② Calculate the frequencies expected by chance.

③ Calculate χ^2 test statistic from the observed & expected frequencies

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- f_o = observed frequency for that cell
- f_e = expected frequency for that cell.

④ Compare it against critical values from a stats table.

- Contigency Table.

careers	observed frequency	expected frequency	observed - expected
research	10	14	-4
teaching	5	14	-9
clinical	30	14	16
occupational	15	14	1
educational	10	14	4
total	(70)		

(Q83) What are the odds of being a man if you watch sport?
No of men who watch sport / No of Women...

$$\chi^2 = \frac{(-4)^2}{14} + \frac{(-9)^2}{14} + \frac{(16)^2}{14} + \frac{(1)^2}{14} + \frac{(-4)^2}{14}$$

$$= (26.4)$$

Chi-square critical value table: - we can reject Ho when: χ^2 observed $\geq \chi^2$ critical.

df	$p \leq 0.10$	$p \leq 0.05$	$p \leq 0.01$	$p \leq 0.001$
1		3.84		
2		5.99		
3		7.81		
4		9.49		
5		11.07		

df = number of categories - 1

∴ critical value = 9.49
since 26.4 > 9.49 we can reject the null hypothesis.

∴ The students do not randomly chose careers. This distribution is significantly different to that expected by chance.
($\chi^2(4) = 26.9$, $p < 0.05$)

multiple i.v.s

The only real difference in the test with multiple i.v.s is calculating how many observations are expected:

$$\text{expected} = \frac{\text{row total} * \text{column total}}{\text{grand total}}$$

$$\chi^2 = \sum \frac{(f_o - fe)^2}{fe}$$ the same

↳ $\chi^2 = \frac{(6-5.85)^2}{5.85} + \frac{(9-9.15)^2}{9.15} + \frac{(13-13.65)^2}{13.65} \ldots$

$= 3.44$ (males + females)

observed	male	female	
occup	6	9	15
teaching	13	22	35
educat	5	5	10
clinical	9	21	30
research	6	4	10
	39	61	100

expected	male	female	
occup	5.85	9.15	15
teaching	13.65	21.35	35
educat	3.9	6.1	10
clinical	11.7	18.3	30
research	3.9	6.1	10
	39	61	100

- degrees of freedom (for multiple i.v.s)

$df = (\text{rows} - 1)(\text{columns} - 1)$

→ 5 × 2 categories so $df = (5-1) \times (2-1) = 4$ therefore critical value = 9.49

- Since $3.44 < 9.49$ we cannot reject the null.
- No individual cell(s) in the matrix had a particularly surprising frequency.

SPSS χ^2

- for a single i.v.

  Analyze / Non-parametric Tests / Chi-square

- for multiple i.v.s.

  Analyze / Descriptive Statistics / Crosstabs

Binominal Test

another test used for data that can only have two values.

- Like χ^2, it compares whether the f of occurance for two values has some expected ratio.

eg. coin → 50% 'heads/tails or not?

Analyze / Non-Parametric Tests / Binominal test

↳ A binominal test indicated that there were more females on the course than would be expected by chance alone ($nF = 61, nM = 39$, $p < 0.05$)

χ^2 assumptions

- from SPSS help pages:
  - non-parametric tests dont require assumptions about the shape of the underlying distribution
  - the data are assumed to be a random sample
  - the expected frequencies for each category should be at least 1. No more than 20% of the categories should have expected frequencies of less than 5.

- low frequencies
  - To deal with low frequencies you can:
    - collapse categories (vale to 1 sto allo)
    - discard small category

eg "research" category can't be combined with any other category so we exclude it by filtering the data

Analysing 3 or more Conditions Part I

LECTURE 11

- issues with multiple t-tests

eg. Are there differences in the abilities of three groups (psychologists, sociologists and astrophysists) to guess a person's mood?

→ we could run 3 tests psych vs soc
soc vs astro
psych vs astro

BUT:

problem: if each test has an alpha of 0.05 the overall likelihood of error becomes (just under) 0.15 which is too high.

} we would decrease alpha to 1.666%. but this ↓ power !!

ANALYSIS OF VARIANCE

is an extension of a t-test
it allows us to test whether 3 (or more) population means are the same, without reducing power.

ANOVA → analysis of variance

VERSIONS:

① one way or one factor ANOVA
multi-factor ANOVA (2x3 anova)

- is a way to compare multiple conditions in a single, powerful test
- it compares the amount of variance explained by our experiment with the variance that is unexplained.

② MANOVA → extension of ANOVA w/ multiple d.v.s
multivariate analysis of variance

③ ANCOVA → extension of ANOVA for continuous variables eg. correlations
analysis of covariance

ASSUMPTIONS:

- scores were sampled randomly and are independently
- roughly normal distribution
- roughly equal number of participants in the groups (improved power if exactly the same)
- roughly equal variance for each condition

Between Groups ANOVA

- the explained variance is the variance between groups.
- the unexplained variance is the variance within a group.

remember: $\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{N}$

ANOVA is based on the F-ratio

$$F = \frac{MS\ between}{MS\ within}$$

→ the effect
→ the noise 'error'

MS = means squared.

Q. SS = a sum of squares divided by its degrees of freedom

- degrees of freedom

ANOVA critical values require 2 d.f. values,
one for ⊙ df between conditions
⊙ residual degrees of freedom

we have to report both.

Q.50 ⊛ $MS_{Between} = \frac{Sum\ of\ Squares}{df}$ 79.86

⊛ $MS\ Within = \frac{Sum\ of\ Squares}{df}$

$\frac{MS\ Between}{MS\ within}$

Pair-Wise Comparrisons

- ANOVA tells us whether groups differ or not
- How do we know which particular conditions?
  - ↳ run the multiple comparisons (those which we were trying to avoid)
  - ↳ some of these are 'planned comparisons'
- some are 'post-hoc' tests
  - ↳ correct for multiple comparisons

example ⊕ Q.63

example:

- objective - test 'social skills' of psych, soc, astro
- experiment - give each group a series of photographs of people and ask for happy/sad face
- correct percentage score

SPSS

- single I.V.
  Analyze / Compare Means / One-Way-Anova
- general approach
  Analyze / General Linear Model / Univariate

I.v is a **Fixed factor**

Terminology

A (number of i.vs) -way (how these variables were measured) ANOVA:

eg. one-way independent ANOVA
= 1 i.v. measured using different entities.

eg. two-way repeated measures ANOVA
= 2 i.v. both measured using the same entities

(Q.60.) (samples + size from tables)

degrees of freedom for between samples
df = 4 ~ 4+1 = 5 levels.
(d.f = N-1)

Total N of participants = 44
(N-1) ∴ ~ 44+1 = 45. 45 people split into 5 groups = 9 per group.
∴ **5 samples of size 9.** answer

Summary

① Only run/report post-hoc tests if main test found a significant result!

② Great many choices, heavily debated area

③ Bonferroni's method:
- uses a t-test and then divides alpha by some value based on the number of tests
- very strict: safe but not v. powerful

Glossary

- unexplained variance (within) = residual or error
- one-way = one independent variable
- univariate or (multivariate) = one dependent variable
- SS = Sum of Squares
- MS = Mean Square = $\frac{\text{Sum of Squares}}{df}$

Possible Outcomes

after running the anova, we might want to compare 2 of the 3+ variables eg. start & end w/out the in-between. We conduct an ANOVA on the whole set and then a single, planned t-test on comparing the 2

Analysis of Variance Part II

LECTURE 12

- Between group Vs. Repeated Measures →

* For repeated-measure designs, there are three possible sources of variance:

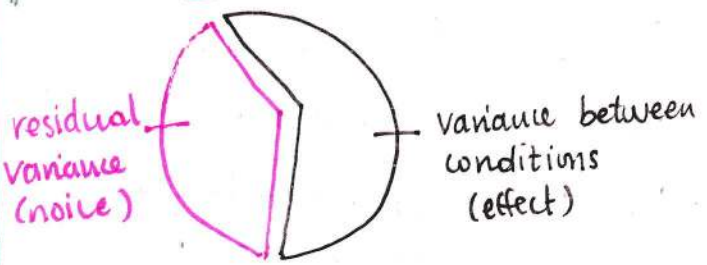
① Variance between conditions

② Variance between subjects (individual differences)

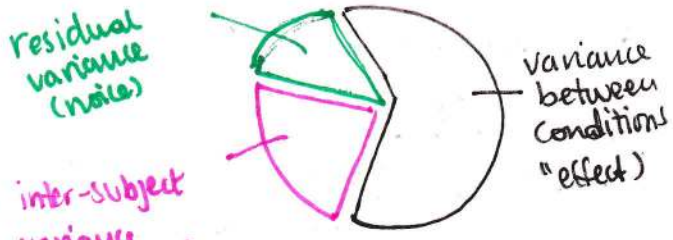
③ residual (unexplained) variance

* In the between-group study, the variance between subjects fell under the category 'unexplained' (see →)

BETWEEN GROUP



REPEATED-MEASURES



→ In both cases, we want to know about the ~~effect~~ ratio of the effect to noise (the brown area divided by the residual variance)

→ The ratio is larger after removing the inter-subject variance.

F-Ratio

as before, we compute the F-ratio and compare it against the critical value.

$$F = \frac{MS\ effect}{MS\ noise}$$

MS effect → variance between conditions

MS noise → remaining variance after accounting for individual differences

$$MS\ noise = MS\ Total - MS\ effect - MS\ ind.diff.$$

MS Total → variance of all subjects within a condition

MS ind.diff. → variance between subjects within a condition

Example:

- objective: We want to know how well children transfer learning in a task
- experiment: create 3 versions (A, B, C) of a problem-solving task and see how long they take to solve it
- measures:
  - Time taken in 1st task as a base line
  - time taken to solve after explaining the task (teaching phase)
  - time taken on 3rd version with no further coaching (transfer)

possible outcomes

run a within-subject ANOVA on the times-to-solve

- no significant difference between conditions = children did not learn
- significant difference caused by short times in the teaching phase = learning.
  significant difference caused by short times in both teaching and transfer phase = good transfer of learning

we mostly care abt the comparison between the baseline and transfer phase. we conduct an ANOVA on the whole, then a single t-test on baseline and transfer

SPSS

Analyze / General Linear Model / Repeated Measures

Multi-Factorial ANOVA

- similar to repeated measures ANOVA
- factors can all be within-subject or all between-group or mixed design
- we can have a 'main' effect or a variety of 'interactions'

- main effect: one of the factors i.v.s consistently affects the d.v. in the same way.
- interactions: the effect of one factor depends on the presence of another

2X2 - ANOVA

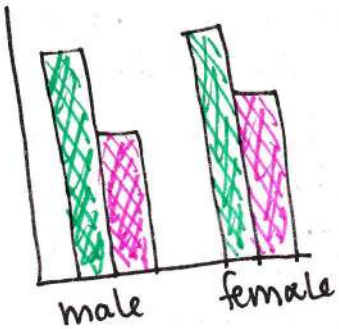
- the multi-factorial ANOVA is a single test
- it returns multiple F values (one for each main effect to be checked and one for the interaction)
- with only 2 levels, there is no need for post-hoc tests eg Q.73: 2x2 repeated measures multifactorial ANOVA would return 3 F-values.
- So 2X2 is just a single test (no family-wise error)
- The equivalent full comparisons using t-test would be to check all combinations of 4 conditions.

More factors: ANOVA would seem a convinient way to take all factors into account at once: it yields an F-ratio and a significante value for any interactions you may be interested in
BUT:
interpreting a 4-way interaction with 3 levels in each factor is hard. There is a trade off between being realistic & being able to interprete the data

2X2 data examples

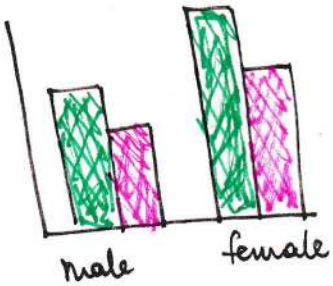
experiment: collect % correct for males and females on a spatial and verbal task. Various types of result: spatial verbal

① Main Effect on Task, No effect of Gender:

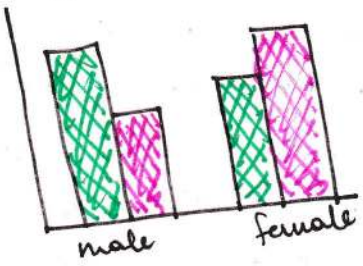


② Main Effect on Gender, No Effect on Task

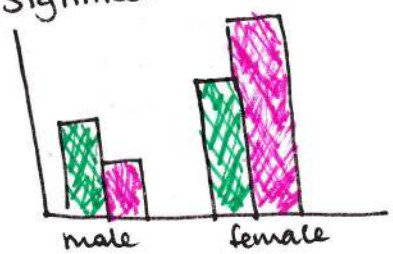
③ Main Effect on Task, Main Effect on Gender



④ No Main Effect of Task No Main Effect on Gender - Significant Interaction



⑤ No Main Effect on Task, Main Effect on Gender - Significant Interaction



Regression Analysis

LECTURE 13

a technique to fit a model (math equation) to data

model can be very simple (few parameters) or complex (many parameters)

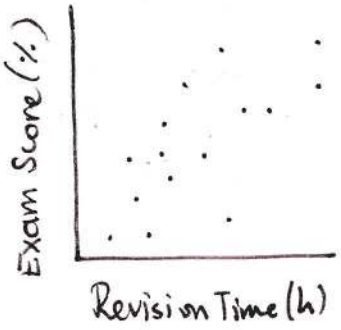
$y = a + bx$

- Criterion or dependent variable
- regression coefficients

→ we can know the slope of a relationship between two variables (gradient)

we can predict y from a given value of x (extrapolation/interpretation)

eg: what is the effect of more revision on grade improvement?



SPSS

Analyze / Regression / Linear

need to know the i.v. & d.v.

. . . . . .

- SPSS output
  - model summary = (r and r^2)
  - coefficients = that's what we are after
  - residuals, variables, ANOVA = ignore

Model	Unstandardise Coefficients		Standardise Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	30.699	4.604		6.668	0.000	21.197	40.200
Revision Time	1.422	.292	.706	4.877	0.000	.820	2.023

We can predict exam % from revision hours

Score = 30.7 + 1.4 * revision hour

at least: score 21.2 + 0.8 revision hour

at most: score 40.2 + 2.0 revision hours.

$= h = (score^{*} - 30.7)/1.4$

If one wants to get 50 %* and spend the rest of the time in a pub one should revise for $h = (50 - 30.7)/1.4 = 4.8$ hrs.

Regression & Correlation

↓

- tells what the relationship is
- doesn't tell whether it is real

→

- show that there is a relationship between the variables and how much noise contributes
- doesn't say much about what the relationship is

Multiple Regression

- simple linear regression tells us about
  - the effect of 1 variable on another
  - the line of best fit to the data
- when we have multiple i.v.s we can look at
  - how they each contribute to the d.v.
  - The PLANE of best fit
- Very similar equation (2 i.v.s)

$$z = ax + by + c$$

eg. Exam score based
- revision hrs
- IQ

Analyze / Regression / Linear...

d.v = Exam %

i.v.s = ① Time revising hrs
② IQ

Summary

- like ANOVA, may hard to portray the data
  hard to visualise on 2D
- excellent way to begin making simple quantitative models and fitting them to your data.

Glossary

- criterion / dependent variable = The variable being predicted.
- predictor variable = a variable used in regression to predict another variable. (sometimes refered as the independent variable)

Non-Parametric Tests

LECTURE 14

• Interval scales and skews

- most data recorded on a ratio or interval scale are normally distributed
- when data are heavily skewed, the step size between points on the scale is probably not constant
- at the simplest level, one should probably be using a non-parametric test for ordinal data

	parametric	non-parametric
correlations	Pearson's r	Spearman's R
2-samples	Paired-samples t	Wilcoxon T
	Unpaired-samples t	Mann-Whitney U
k-samples	Rep-measures ANOVA	Friedman's
	Between-Groups ANOVA	Kruskal-Wallis

example: compare two groups of subjects on learning lists of words and pictures

- rank all scores
- if the null hypothesis is true, the rank totals should be the same.
- if the null hypothesis is false, the rank totals should be different

	Score	Rank
pic	10	5
pic	12	7
pic	16	11
pic	15	10
pic	17	12
pic	14	9
word	6	1
word	7	2
word	8	3
word	11	6
word	9	4
word	13	8

Total = 54

Total = 24

Mann-Whitney U test

- non parametric
- designed to look at differences between treatments in a (between-groups) design
- used when there are only 2 treatment levels (1 i.v. with exactly 2 levels)
- requires d.v. to be on at least an ordinal scale

$$U = n_1 n_2 + \frac{n_x(n_x+1)}{2} - T_x$$

n_1 = number of subjects in group 1

n_2 = number of subjects in group 2

n_x = " " in larger rank total

T_x = total rank (the larger one)

we then compare the observed value of the test statistic U to a critical value: $U_{observed} <= U_{critical}$ →

- when the null hypothesis is true, we would expect the scores of the two groups to be randomly distributed

group	A	B	A	A	B	B	A	B	B	A
score	1	2	3	4	5	6	7	8	9	10

- when the null hypothesis is false, we would expect the scores of the two groups to be clustered at either end of the sequence.

group	A	A	A	B	A	A	B	B	B	B
score	1	2	3	4	5	6	7	8	9	10

in our example: $U = n_1 n_2 + \frac{n_x(n_x+1)}{2} - T_x$

$U = (6)(6) + \frac{6(6+1)}{2} - 54 = 3$

$U_{critical}\ (n = 6,6) = 5$ (?) This value is found from tables ü

since $U_{observed} <= U_{critical}$, we reject the null hypothesis and conclude that subjects can recall more pics than words

this is because the $T_x = 54$ (larger) which is the total of the pics

SPSS → Analyze / Non Parametric Test / 2 Independent Samples...

Wilcoxon Related-Samples test

NOT Wilcoxon Rank Sum Test!

- non parametric - an ordinal scale for the d.v.
- designed to look at differences between treatments in a within subject design
- used when there are only two treatment levels.

Wilcoxon test / rationale

- test is based on the difference between the two scores for each subject
- uses both direction (which score is greater) and magnitude of the differences (how much greater)

→ if H0 is true, then differences in one direction will be as large as the differences in the other.

- that is what the test statistic will measure
- but it will use ranks of magnitudes rather than actual magnitudes so it can be used on skewed data.

example: case I

RT cong	RT incong	diff	absolute diff NEG	absolute diff POS	ranks NEG	ranks POS
602	898	296		296		10
461	494	33		33		2
626	869	243		243		9
402	392	-10	10		1	
439	551	62		62		3
562	635	73		73		* 4.5
514	587	73		73		* 4.5
997	1078	81		81		6
921	1052	131		131		7
1003	1208	205		205		8

$T- = 1$ $T+ = 54$

- the test statistic we are interested in is the smaller T value (whether - or +).
- as with other tests then we compare the observed value of the test statistic, W, to a critical value, found in the tables ü

$W_{observed} < W_{critical}$

↳ we conduct a Wilcoxon paired-samples test and found that the subjects were significantly slower to respond in the inconsistent condition: ($W = 1$, $N = 10$, $p < 0.01$)

* it should be 4,5 but since there is a tie we add $4+5 = 9$ $9 \div 2 = 4.5$

* eg. 32 1
51 2
63 4
63 4
63 4
98 6

$\frac{3+4+5}{3} = 4$

- zero difference scores:

we may get a difference of zero (same in cong/incongr). All those subjects essentially get removed. - their score doesn't contribute
- simply reduce N by the number of subjects where this happens

Analyze/Non Parametric Tests/2 Related Samples

Non-Parametric Tests Part II

LECTURE 15

Kruskal-Wallis H Test

- non parametric
- designed to look at differences between treatments in a {between groups} design.
- used when there are multiple levels (1 i.v. with more than 2 levels)
- requires d.v. to be an ordinal scale

- Kruskal-Wallis H test Irationale

→ When the null hypothesis is true, we expect a random distribution of ranks across groups:

eg:

Gp	A	B	C	B	A	C	C	B	A
Rank	1	2	3	4	5	6	7	8	9

Group	Rank Sum
A	15
B	14
C	16

average ranks of levels in the design should be equal.

→ When the null hypothesis is false, we expect a systematic distribution of ranks across groups:

eg

Gp	A	A	B	A	B	C	B	C	C
Rank	1	2	3	4	5	6	7	8	9

Group	Rank Sum
A	7
B	15
C	23

average ranks of levels in the design should be different.

Q.108 R^2 = The proportion of variance in the outcome accounted for the predictor variable.

- equation:

$$H = \left[\frac{12}{N(N+1)} \sum_i \frac{Ri^2}{ni}\right] - 3(N+1)$$

N = total number of scores

Ri = sum of ranks for each level of variable

ni = the number of the scores in each level of the variable.

$$H = \left[\frac{12}{18 \times 19} \times \left(\frac{66.5^2}{6} + \frac{78.5^2}{6} + \frac{26^2}{6}\right)\right] - 3 \cdot (18+1)$$

H ↓ total number of subjects

↳ nº of subjects

Analyze / Non Parametric Tests / k independent samples

→ SPSS

In SPSS the Kruskal-Wallis test doesn't produce a H. Instead it converts it to a Chi-Square

$\chi^2 = 8.99, df = 2, p < 0.05$

"There was a significant difference between the reinforcement conditions as measured by a Kruskal-Wallis test

TESTING DIFFERENCES →

Q.96 We can tell that the data are not normally distributed by looking at the significance (table OUTPUT) of the K-S (0.005) and Shapiro-Wilk test (0.01). They are both less than 0.05. Hence a non-parametic test is more appropriate.

Q.97 Another term for Non-Parametric test is: Distribution free Test

example (rank across subjects)

Control Score	Rank	Praise Score	Rank	Criticism Score	Rank
10	4	12	6.5	11	5
13	8.5	14	11.5	9	2.5
14	11.5	15	14.5	12	6.5
15	14.5	14	11.5	9	2.5
14	11.5	16	16.5	8	1
16	16.5	17	18	13	8.5
Total = 66.5		Total = 78.5		Total = 26	

- Kruskal-Wallis only tells us that the 3 conditions differ but not which one.

⊙ We can use multiple Mann-Whitney tests and apply the Bonferroni correlation (multiply the p value by the possible number of tests)

experiment

Subject	Control Score	Rank	Praise Score	Rank	Criticism Score	Rank
S_1	10	1	12	3	11	2
S_2	13	2	14	3	9	1
S_3	14	2	15	3	12	1
S_4	15	3	14	2	9	1
S_5	14	2	16	3	8	1
S_6	16	2	17	3	13	1
	Total=12		Total=17		Total=7	

Friedman Test

- non-parametric
- designed for paired samples in both within and matched-subject design
- used when there are multiple levels (1 i.v. with more than 2 levels)
- requires d.v. to be an ordinal scale

rationale

- each individual subject's scores are ranked.

* When the Null hypothesis is true, we expect a random distribution of ranks across the subjects:
rank totals of d.v. for the different levels of the i.v. will be about equal.

Subject	Level A	Level B	Level C
S_1	1	2	3
S_2	2	1	3
S_3	3	2	1
Rank Total	6	5	7

* When the null hypothesis is false, we expect a systematic distribution of ranks across the subjects.
rank totals of d.v. for the different levels of the i.v. will be different.

Subject	A	B	C
S_1	1	2	3
S_2	2	1	3
S_3	1	2	3
Rank Total	4	5	9

Equation:

$$\chi^2_R = \left[\frac{12}{Nk(k+1)} \sum R^2\right] - 3N(k+1)$$

N = total no. of subjects per group
R = rank total for each level of the I.V.s
k = number of levels in the I.V.

$$\chi^2_R = \left[\frac{12}{6 \cdot 3 \cdot (3+1)} \cdot 12^2 + 17^2 + 7^2\right] - 3 \cdot 6 \cdot (3+1)$$

Analyze/Non Parametric Tests/k Related Samples

again reported as χ^2

testing differences:

- as with Kruskall-Wallis, the Friedman test does not tell us which conditions differ:
  - it can be dealt using a multiple bonferroni-corrected Wilcoxon test.

Resampling Statistics Part I

LECTURE 16

Resampling techniques represent a novel method that is assumption-free (er) but retains power.

MOTIVATION – Why use resampling techniques?

- fewer assumptions
  - more accurate (where assumptions had not been met)
- very general
  - few basic ideas, that can be modified and reused.
- no equations or tables to look up
- they force us to think about our data

Why are resampling approaches not more popular?

- new
- reasonably good job, and are discussed in simple (!) language in textbooks
- resampling does require some programming (not SPSS)
- there is a perception that new techniques must be more complex.

- to determine the probability that the differences we measured were caused by sampling error (motivation for inferential statistics).

→ the principle of resampling techniques is to measure that sampling error by repeating the sampling process a large number of times:

we can determine the likely error introduced by the sampling by looking at the variability in the resampling.

within-subject randomisation test

H0: taking steroids has no effect on a dwarf's height.

- Now the population that we randomize are within-subjects: in each resample, the values are shuffled for each subject rather than across the whole dataset: We just randomise the sign of the difference for each pair.

between-subject randomisation tests (a.k.a. permutation tests)

Q: are smurfs just dwarfs painted blue?

Exp: measure heights & check for difference

H0: smurfs & dwarfs have the same height.

Smurfs mean height: 28.1

Dwarfs mean height: 32.8

→ we want to determine the likelihood of getting differences this extreme if the data all came from a single population.

S	D
30	35.2
31.5	15.2
23	⋮
⋮	

mean = 28.1 | 32.8

difference = 4.7

randomise / mix them up →

S	D
35.2	49
31.5	24.2
⋮	⋮

mean = 26.0 | 34.9

difference = 8.9

repeat many times

- we assume that these are real & sensible values
- we don't assume anything about their distribution.

- our measured difference between means = 4.7
- 3304 resampled differences were more extreme than 4.7, which is 33% of 10000.
- so 33% of values were more extreme

$p = 0.33$ (> 0.05)

(repeat many times, assuming H0 is true, and check how extreme the real value was.)

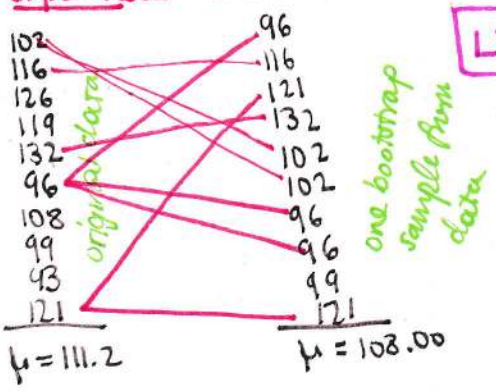
Resampling Statistics Part II

LECTURE 17

There might not always be as many as 1000 possible ways to shuffle the original sample:
→ more efficient to use only the possible resamples.

- for between-group design with N subjects per group $(2N)!/N!^2$
- for repeated measures 2^N

experiment: measure IQ of 10 smurfs



bootstrap Vs Resampling

resampling techniques measure the sampling error by repeating the sampling process many times.
→ we can determine the likely error introduced by the sampling by looking at the variability of the resampling.

resampling test	bootstrap
compare 2 different conditions	work with single group
shuffle dataset to create resampled datasets: Randomise whether subjects had been 'in' condition A or B	"resample with replacement"

USAGES & PROPERTIES

- Bootstrap resamples can be used to calculate
  - confidence interval of a mean eg 95%.
  - standard error of mean
- Determine whether some test value is inside or outside the 95% confidence interval (like a one-sample test).
- These are non-parametric methods, making very few assumptions of the data.

* calculating a s.e.m.

- the s.e.m is the standard deviation of the means of all possible samples.
- it can be estimated from the standard deviation of the bootstraps.

in our example:
s.e.m. is based on the formula
s.e.m. based on the bootstrap resamples.
The difference is due to the fact that they are both ESTIMATES, calculated in 2 different ways.

* calculating a confidence interval
the 95% of the confidence interval from the bootstraps represents the range of values that 95% of the means take.
→ calculate from bootstrap means by ordering them and cutting off the highest/lowest 2.5%

07 92380419

a one-samples-test

Consider that the hypothesis of this population is "above average IQ". Then we might expect that IQ > 100

H_0: IQ ≤ 100

i.e. we want to know whether the mean IQ is significantly greater than 100.

- order the data and find how many values are ≤ to 100.
- in our example: 45 values ≤ 100

so $p = 45/10000 = 0.0045 < 0.05$

bootstrapping confidence Intervals

① Take bootstrap resamples of the subjects (i.e. each subject's data might be included in the analysis once, more than once, not at all).

② For each resample, perform the analysis (eg. measure the line of best fit).

③ Percentiles (or some other measure of spread) can be taken directly from the bootstrap distribution of the statistic

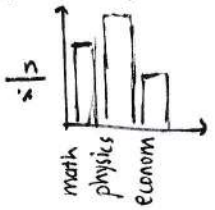
Graphing Data

LECTURE 18

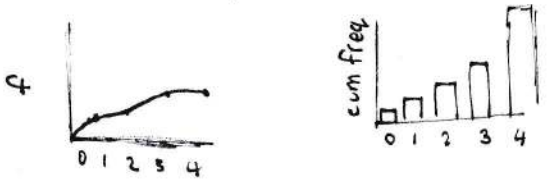
• HISTOGRAMS

frequencies:

histograms and (percentage histograms) for categorical data



histograms and (cummulative histograms) for binned numerical data



Graphs / Legacy / Histogram

• ERROR BAR PLOT

- indicate the variability or confidence interval for a data point
- standard deviation, s.e.m, c.i.s can all be used to make error bars

Variance and error bar plots

→ often data vary in multiple ways: occasions, multiple subjects, multiple stimuli: To which source of variability should we associate an error bar?

→ often at inter subject variability

Graphs / Legacy / Error Bar

- simple: only if you want to consider one i.v.
- clustered: if you want to consider two i.v.s.
- summarises for groups of cases: between subject design
- Summarises of separate variables: within subject design.

• SCATTER PLOT

- useful when there are several levels of some i.v.

eg. Sternberg data for each subject with measurement of RT for multiple angles.

- adding a line of best fit

double click graph / Add chart Element / Fit Line at total

- overlaying additional data

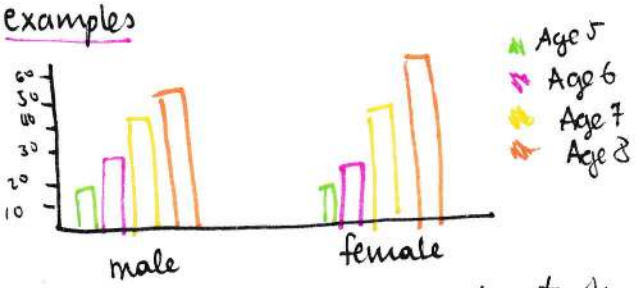
Transform / Compute /

• MULTIDIMENTIONAL DATA

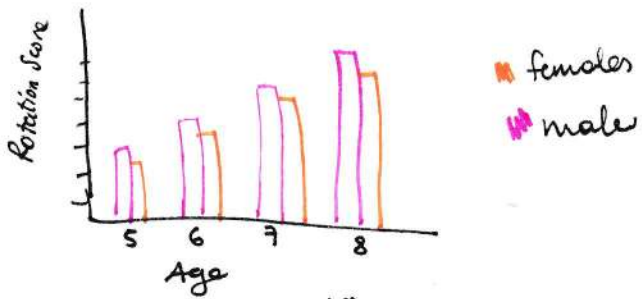
→ more than 1 i.v. (most experiments)

eg. a two-factor (made up) experiment with six different age groups and 2 different genders, on a spatial rotation task.

examples



→ Both sexes get better at the task as they grow older.



→ Gender Differences

pseudo-colour

- colour is a popular way to show intensity, it can potentially introduce effects that are not really there.

- 4 Dimensions

eg. 3 ivs with 1 d.v.

Bringing it all Together

LECTURE 19

• Supporting a Theory

- common scientific pursuit
- typically, a scientist tries to find evidence to support his theory
- This is not the best way - impossible

• comparing two theories

- Theory A predicts X, whereas theory B predicts Y.

• disproving a theory

- when scientist A runs experiments that rule out scientists' B's theory, scientist B can modify / refine their theory to account the new data.

NEGATIVE RESULTS

→ what happens when you find no significant difference between 2 conditions?

The chosen test only showed that you have no reason to reject the null hypothesis.

That doesn't necessarily mean that the Null hypothesis is necessarily true.

→ Negative Data can be harder to be interpreted

TESTING FOR SIMILARITY

- There is only a test that looks only for significant <u>differences</u>.

Research hypothesis → statistical hypothesis

↳ the grand theory leads to a more specific research hypothesis

↳ Chimpanzee recognises himself in the mirror.
+ presence of a spot on his forehead:

i.v & d.v. relate

- H0: Chim <u>will not</u> touch his forehead.
- H1: Chim will touch his forehead

3 Challenges:

- which test to run?
- what are the test's requirements?
- how to perform test in SPSS

Revision

LECTURE 20

• Transform & z-scores

to tranform data is to apply a mathematical operation to each individual value.

A way of transforming data from individual subjects onto a consistent, universal unit, namely "number of std devs from the mean".

Simple formula: subtract the mean from each score and divide by the staund. dev.

z-score: +0.5 means that participant was 0.5 std devs above the mean

Because of what we know about staund. dev. roughly 68% of subjects z scores will be in range -1 to 1. etc.

So, from a z-score we know how good/bad a particular score was on a standard scale

• Distributions

- kurtosis → platykurtic
  → Leptokurtic

• parametric tests

how to assess deviation from normal distribution?

- plot a histogram that shows both raw data and a normal curve: does curve capture data?
- ran a test for normality
- simplest level: if data are on an interval/ratio scale, use then parametric tests.

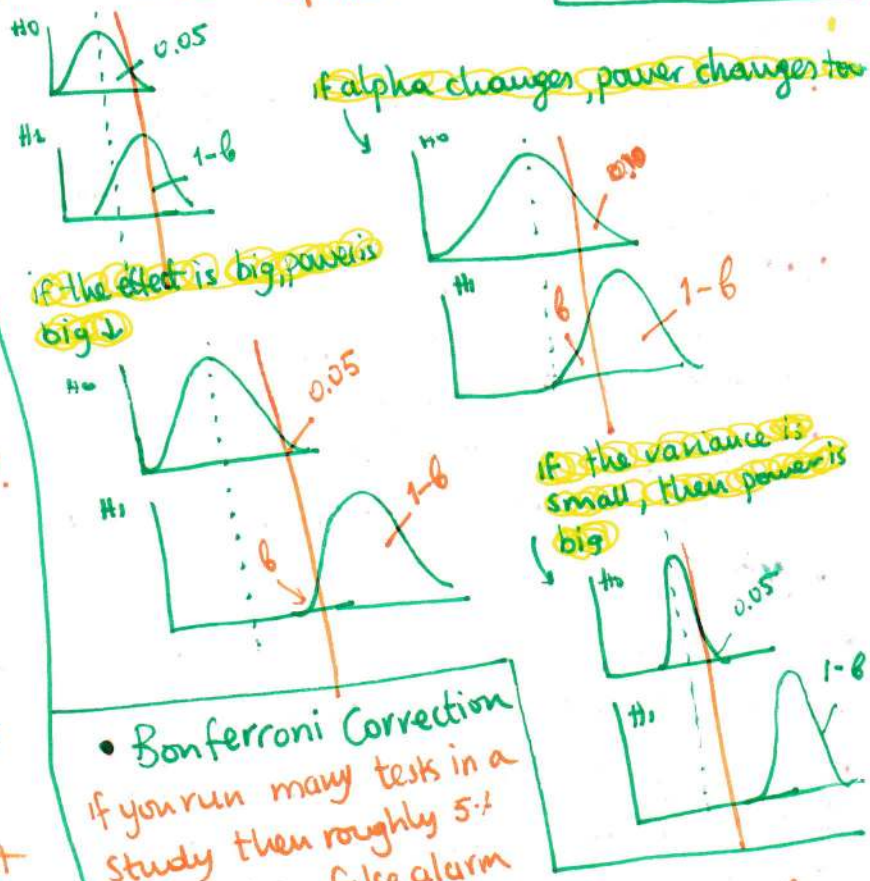
• Hypothesis Testing

H0: assume that our variables have no effect

H1: unless that seems really unlikely

if the probability is low (less than alpha) then it is improbable that there is no effect, therefore we have found evidence that there is an effect.

• power and alpha



• Bonferroni Correction

if you run many tests in a study then roughly 5% will give false alarm

↳ REDUCE YOUR ALPHA!

- divide alpha, your criterion value for p by the number of tests
- multiply the p value itself by the number of tests and keep the criterion the same.

• Regression Analysis

- relationship between two variables.
- what that relationship is.

• ANOVA

- unrelated-samples ANOVA
- repeated-samples ANOVA
- what does a main effect of a certain value mean?
- what is an interaction?

• Resampling

uses all data, work out probabilities using simulations rather than formulas, makes fewer assump.

randomisation / permutation

- shuffle the data to simulate the null hypothesis.
- can be used for within/between subject design
- can therefore be used for one-sample test

bootstrapping: • provides confidence intervals or sem for individual conditions

- resampling

randomisation tests

- recreate data, using the original, where you know the null hypothesis is true.
- Count how often in your new data there is an 'effect' as big as your original one: this is p
- if p is less than 0.05 then this unlikely to have occured by chance:

- χ^2

- the test to determine how often observations fall into a particular category.
- χ^2 compares how many observations fall into a category with how many were expected based on some model.

Null Hypothesis: the different frequencies observed for different categories were expected by chance

Alternative Hypothesis: the different frequencies observed for different categories reflect actual differences between categories

EQUATIONS

1. mean = $\frac{\Sigma x}{n}$

2. variance : $\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{N}$

3. standard deviation : $\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{N}}$

   from sample : $s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{N-1}}$

4. Gaussian Equation : $y = peak \times e^{-\left(\frac{x-\mu}{2\sigma^2}\right)}$

   x = desired score
   y = frequency at x
   μ = mean of pop
   σ = standard deviation of pop

   $peak = \frac{N}{\sigma\sqrt{2\pi}}$

5. Skewness : $\frac{3(mean - median)}{standard\ deviation}$
   ↳ removes influence of variability

6. z-score : $z = \frac{score - \bar{x}}{\sigma}$

7. s.e.m. : $\frac{\hat{\sigma}}{\sqrt{N}}$

8. confidence Intervals : $\bar{X} \pm 1.96 \times s.e.m.$
   95

9. t = mean difference of 2 conditions divided by the standard error difference of the 2 conditions

   a) (mean of GP1) − (mean of GP2)

   b) s.e.m. $(\bar{X}_1 - \bar{X}_2)$: $\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$

   $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$

10. Covariance : $\frac{\Sigma(\bar{x}-x)(\bar{y}-y)}{N-1}$

11. Chi-Square: $\chi^2 = \sum \frac{(fo-fe)^2}{fe}$

- fo = observed frequency
- fe = expected frequency

for multiple i.vs fe: $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$

degrees of freedom for multiple i.v.s : $df = (\text{rows}-1)(\text{columns}-1)$

12. F-ratio: $F = \frac{MS\ between}{MS\ within}$ → the effect; → the noice 'error'

- MSBetween = $\frac{\text{Sum of Squares}}{df}$
- MSwithin = $\frac{\text{Sum of Squares}}{df}$

↑
- between-groups
- within-groups $F = \frac{MS\ effect}{MS\ noice}$ → variance between conditions

$MSnoice = MStotal - MSeffect - MS_{indiv.\ differ.}$

→ remaining variance after accounting for individual differences

13. $U = n_1 n_2 + \frac{n_x(n_x+1)}{2} - T_x$

n_1 = N of subjects in GP1
n_2 = N of subjects in GP2
n_x = N of subjects in larger rank total
T_x = Total rank in the large one

14. $H = \left[\frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i}\right] - 3(N+1)$

N = total number of scores
R_i = sum of ranks for each level variable
n_i = the number of the scores in each level of variable

15. $\chi^2_R = \left[\frac{12}{Nk(k+1)} \Sigma R^2\right] - 3N(k+1)$

N = total number of subjects per group
R = Rank total for each level of the i.v.s
k = number of levels in the i.v.

NON-PARAMETRIC TESTS

① Mann-Whitney U test (Ranks!)

- non parametric
- designed to look at differences between treatments in a between-groups design
- used when there are only 2 treatment levels (1 i.v. with exactly 2 levels)
- dv needs to be on at least ordinal scale
- U observed ≤ U critical, we reject the null hypothesis
- when null hypothesis is false, the scores of two groups are distributed at either end of the sequence
- if the null hypothesis is false, the rank totals should be different

② Wilcoxon Test (ranks!)

- non-parametric
- ordinal scale for the d.v.
- designed to look at differences between treatments in a within-subject design
- used when there are only two treatment levels (1 i.v. and 2 levels exactly).
- uses both direction (which score is greater) and magnitude (how much greater).
- if H0 is true, then differences in one direction will be as large as the differences in the other
- We are interested in the small T: Wobserved < Wcritical we reject the Null hypothesis

③ Kruskal-Wallis H test

- non-parametric
- designed to look at differences between treatments in a between-groups design
- used when there are multiple levels (1 i.v. with 2+ levels)
- dv. must be an ordinal scale
- when the null hypothesis is false, we expect a systematic distribution of ranks across groups, and average ranks of levels should be different
- Tells us that conditions are different, but not which one (can be dealt using Bonferoni correlation using multiple Mann-Whitney

④ Friedman Test

- non-parametric
- designed for paired-samples in both within and matched-subject design
- used when there are multiple levels
- d.v. must be ordinal scale
- When the null hypothesis is false we expect a systematic distribution of ranks, and different rank totals
- Tells us that conditions are different but not which one. It can be dealt using a multiple Bonferroni-correlated Wilcoxon test